Automating Front-End SoC Design With NetSpeed's On-Chip Network IP

By Tom R. Halfhill
Senior Analyst

March 2015

The Linley Group

www.linleygroup.com

Automating Front-End SoC Design With NetSpeed's On-Chip Network IP

By Tom R. Halfhill, Senior Analyst, The Linley Group

This white paper describes NetSpeed's NocStudio design tool and the Orion and Gemini licensable network-on-chip (NoC) products. This paper was prepared by The Linley Group and sponsored by NetSpeed, but the opinions and analysis are those of the author.

NetSpeed Systems, a three-year-old Silicon Valley startup, is known primarily as a licensable network-on-chip (NoC) vendor, but that's only part of the story. The company's larger ambition is to automate much of the system-on-chip (SoC) front-end design. Whereas back-end flows have benefitted enormously from automated tools for logic synthesis, intellectual-property (IP) integration, place-and-route floor planning, and silicon verification, the front-end flow has changed relatively little since the 1990s. It's still common for chip architects to estimate a design's performance by typing the specifications and parameters into an Excel spreadsheet. Subtle timing problems often remain undetected by the C-level simulation or even the RTL-level implementation. Eventually, those problems show up late in the tape-out process—or, worst case, when the first silicon is dead on arrival from the fab. And even if the chip works properly, it often misses its performance targets.

As SoCs grow more complex, and as new fabrication processes explode the number of design rules, the risk of failure grows more likely and more costly. Cache coherency among numerous IP blocks is the latest wrinkle that complicates timing and design validation. Some companies have simply given up, resigning themselves to buying less-optimized embedded processors on the merchant market. Others are outsourcing their chip designs, which merely relocates the risk.

NetSpeed's founders and technologists are seasoned chip-design veterans who have felt all these pain points, including the dismay of a canceled project after a disappointing tapeout. They have also experienced the satisfaction of finishing many successful designs. (Co-founder/CEO Sundari Mitra began her career working on Intel's 80286 processor in the early 1980s.) Hence their ambition to streamline the initial phase of designing a complex SoC.

NetSpeed's design-automation tool for chip architects is called NocStudio. It's a logical outgrowth of the company's on-chip interconnect technology. Currently, NetSpeed offers two NoC products as licensable IP: Orion, a configurable on-chip interconnect fabric, and Gemini, which builds on Orion by adding cache coherence for processor cores, acceleration engines, and other components. Customers are using these NoCs for chip designs targeting mobile systems, midrange servers, high-end enterprise networks, and FPGAs.

NocStudio is a pre-RTL design tool that enables architects to customize these NoCs and readily compare alternative topologies before committing the design to a C-level simulation or RTL. NocStudio is also capable of automatically generating a network

topology that connects all the IP blocks in a preliminary layout that optimizes the design for performance, power efficiency, die area, low latency, and deterministic quality of service (QoS). In addition, it's a correct-by-construction design tool that prevents fatal errors such as protocol- and network-level deadlocks.

NocStudio's final output includes performance statistics, the RTL files required to synthesize the NoC, a C++ functional model, and verification test benches. By speeding the design process and reducing risks, NocStudio encourages more-thorough design exploration while cutting costs and shortening the time to market.

Licensable IP vs. Home-Grown Solutions

NetSpeed's main competition is not other NoC vendors but rather the proprietary buses, crossbars, and fabrics that many SoC architects still cobble together for their on-chip interconnects. Industry inertia keeps these legacy solutions alive. Although they are adequate for simple chip designs, it's increasingly difficult for conventional interconnects to meet all the latency requirements while preventing deadlocks and minimizing the masses of wiring that inflate die area and power. In addition, IP blocks are growing more diverse in speeds and I/O traffic, requiring multiple networks or segregated subnetworks to avoid bottlenecks. Coherency adds a new layer of complexity, particularly in leading-edge designs with heterogeneous compute capability.

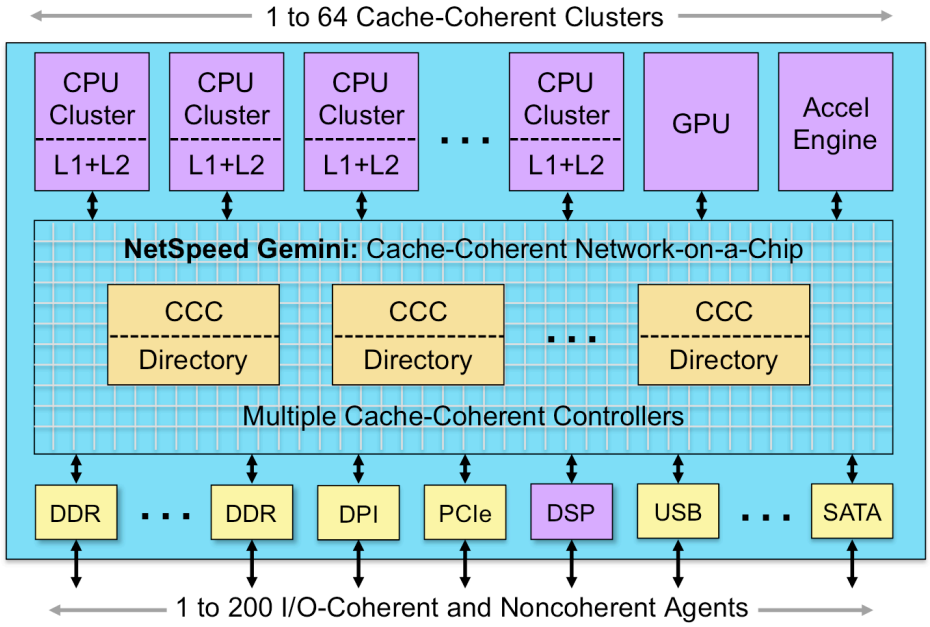

Figure 1. NetSpeed's Gemini network-on-chip. The processing clusters can be homogeneous or heterogeneous, and the directory-based coherence hardware can scale to larger designs than conventional cache snooping.

To overcome those problems, NetSpeed's Orion NoC is designed for chips that don't need cache coherence, whereas the Gemini NoC is designed for chips that have cache-coherent clusters of processor cores (CPUs, GPUs, or DSPs). Typically, all the cores in a cluster share a common L2 cache, and each cluster is a single network node. As Figure 1

shows, Gemini supports up to 64 processor clusters and up to 200 other components that may be I/O coherent. Using quad-core clusters, for instance, Gemini enables a massively parallel chip design with up to 256 CPUs. Alternatively, the design can be heterogeneous, integrating clusters of different processing cores.

NetSpeed applies the industry's experience with macroscale computer and telecommunications networks to on-chip networks. Fundamentally, all networks perform the same functions, whether they span the globe or a tiny chip. To ensure QoS, signals must travel from Point A to Point B within a specified time and without delaying other signals. Therefore, NetSpeed uses the mapping and optimizing algorithms originally developed for much larger-scale networks.

NocStudio can automatically configure an Orion or Gemini NoC. It integrates IP cores from multiple vendors, enables architects to explore alternative designs, and estimates the chip's performance, power consumption, and die area. Although NocStudio is a graphical tool, it doesn't restrict users to a particular design methodology. As the screen shot in Figure 2 shows, two primary design methods are possible:

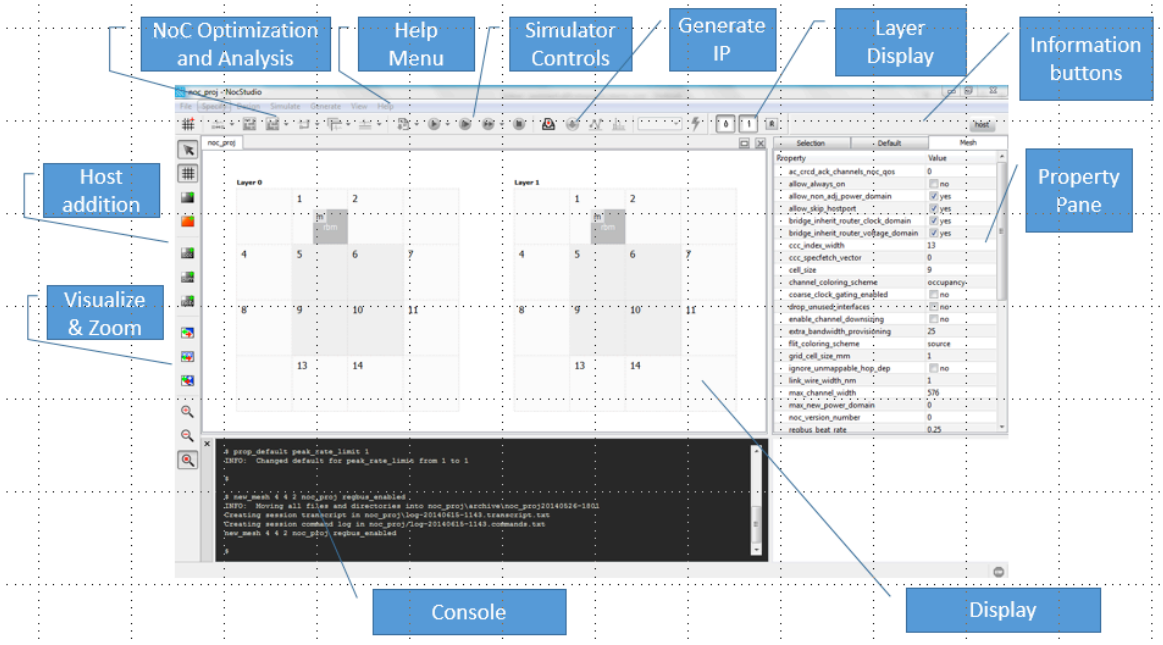

Figure 2. NocStudio screen photo. The main window graphically depicts a high-level view of the chip design; the lower window shows the corresponding definition statements for the synthesis script; and the right-hand window displays property sheets for IP blocks. (Source: NetSpeed)

Using the first method, architects can drag and drop all the desired IP blocks into NocStudio's main window. With each addition or modification, NocStudio automatically displays a script in the lower window that defines the IP blocks for the synthesis compiler. (IP blocks identify their interfaces and other features to NocStudio using the industry-standard XML-based IP-XACT format.) Architects exploring various design options favor this graphical drag-and-drop method; it's also a good introduction to NocStudio.

With the second method, architects can manually write or edit the script that defines the IP blocks by using the command-line interface in the lower window. The architects can also append verbose comments to document their decisions and guide the circuit designers. As the architects add statements to the script, NocStudio automatically updates the design's graphical depiction in the upper window. Customers modifying existing chip designs often favor this method because it leverages their past work. Experienced users can also change the script more quickly than working through the graphical interface.

NocStudio is topology agnostic. The tool will find the best routing solution from among a multidrop bus, ring, tree, mesh, or hybrid topology. Its algorithms try to determine the most efficient mapping and may change the topology as architects add new components or modify the specifications. Alternatively, architects can specify a particular topology, overriding the tool's choice.

As the design evolves, NocStudio updates all the performance statistics, enabling architects to freely experiment with alternatives and make different trade-offs. Those statistics include the link cost (the number of wires required for the interconnects) and the buffer cost (the number of flip-flops required to implement the necessary FIFO buffers). To meet latency requirements and guarantee QoS, NocStudio can automatically add pipeline stages to long wires, or architects can do it manually. The QoS specifications may include such factors as the data-path bandwidth, transfer latency, service priority, and rate limits.

The design process isn't wholly automated—the architects are still active participants. NocStudio simply enables them to rapidly try different network configurations to find the best solution in less time. Of course, architects can do the same thing with spreadsheets and C-level simulations, but optimization is slower and more error-prone, which discourages experimentation.

Optimizing the Chip Design

To optimize traffic among IP blocks that may have different latency, bandwidth, or protocol requirements, NocStudio can vary the data-path widths from 8 to 1,024 bits and create up to 8 heterogeneous physical networks and 32 virtual networks. (Virtual networks appear as separate NoCs but use the same wires.)

Because Orion and Gemini are intended mainly for ARM-based SoCs, they connect directly to IP blocks that support AMBA and AXI protocols. Currently, they support protocols up to AMBA 4; an AMBA 5 version is in the works. NetSpeed or its customers can create gaskets for other protocols. At the network level, Orion and Gemini convert all traffic into a native format called the NetSpeed Streaming Interface Protocol (NSIP).

Figure 3 is another illustration of the NocStudio design process. Architects start by adding IP blocks, which generally have different characteristics. (Processor cores have more logic gates and need greater bandwidth than a UART, for example.) Design rules for the target IC process enable NocStudio to estimate the chip's performance statistics and suggest an approximate floor plan. (The floor planning is high level, however; NocStudio is not a place-and-route tool.) NocStudio will then propose a NoC design that best satisfies the architecture specifications.

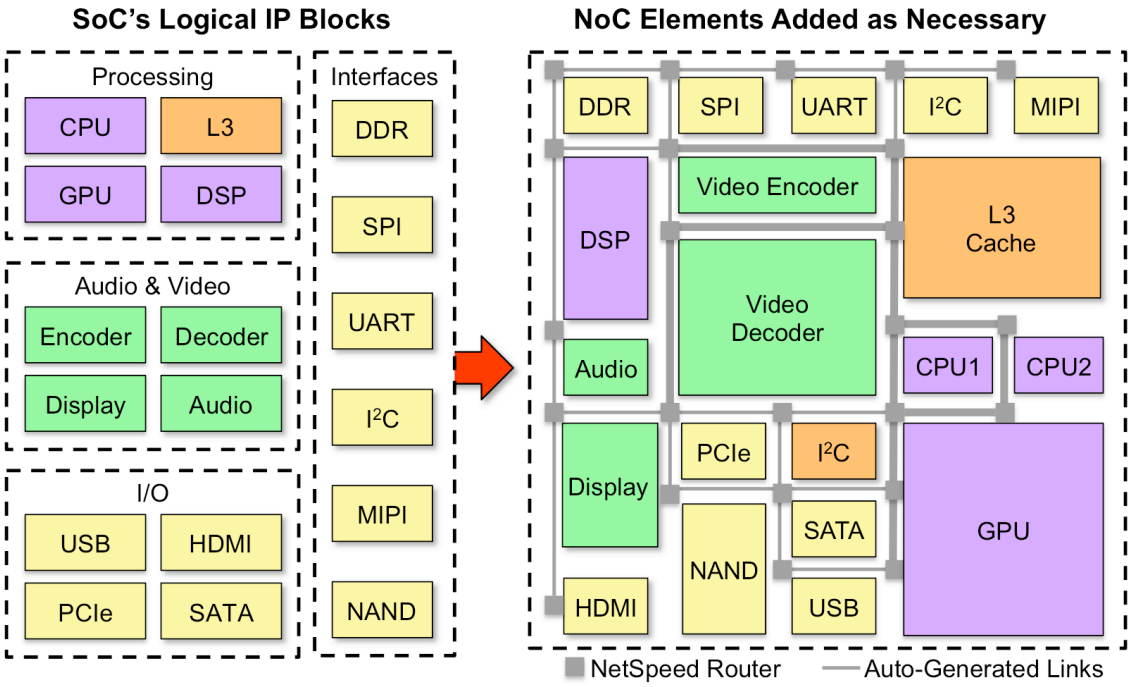

Figure 3. Designing a NoC with NocStudio. This graphical tool can help automate an SoC design using optimal-path algorithms adapted from computer networking and telecommunications. The connecting lines in the right-hand diagram represent the NoC's data paths. Thicker lines indicate wider pathways; users can hover the cursor over a line to display the actual number of wires required. (Source: NetSpeed)

NetSpeed can't foresee which IP cores its customers will license or develop for their chips, so the NoC must accommodate cores with many different characteristics, such as the type of traffic the cores generate and their latency tolerance. To bridge those differences, Orion and Gemini (like other NoCs) are packet-switched networks: they convert all internal traffic into packets that traverse the NoC and are reconstituted into the original bit streams at their destinations. As Figure 4 shows, NetSpeed claims that packetizing the traffic significantly reduces the interconnect wiring compared with a conventional AMBA AXI bus.

At the physical level, NetSpeed's NoC is a grid of point-to-point links between the IP cores. The grid includes not only the interconnects but also switching nodes at the intersections. In effect, these nodes are chip-scale routers that perform the same basic function as any network router—they read the packet headers to steer the payloads toward their destinations. NocStudio automatically creates as many routers as needed, along with the interconnects and appropriately sized packet buffers.

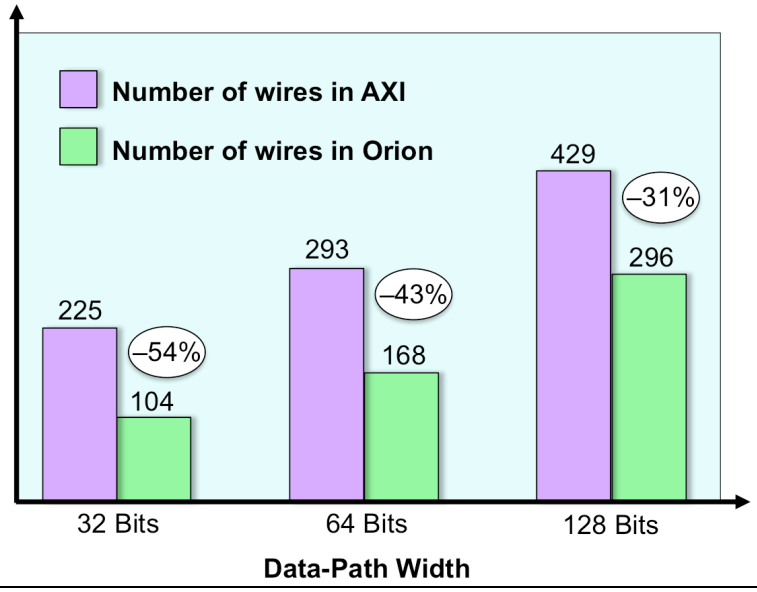

Figure 4. Orion versus AMBA AXI. NetSpeed's packet-switched NoC can reduce the wiring needed to connect the IP cores on a chip. Most consumer and mobile SoCs have data paths in the 32- to 128-bit range. (Source: NetSpeed)

To further reduce the wiring, the routers distribute traffic arbitration throughout the network instead of using a central arbitrator. By optimizing these and other elements, NetSpeed says it can reduce the NoC's power consumption by more than 60% compared with an AMBA AXI bus, as Figure 5 shows. This example assumes a design with 32 master devices, 4 slave devices, and data-path widths ranging from 64 to 256 bits. NetSpeed says the percentage of power savings can grow with design size and complexity because NocStudio can find more elements to optimize.

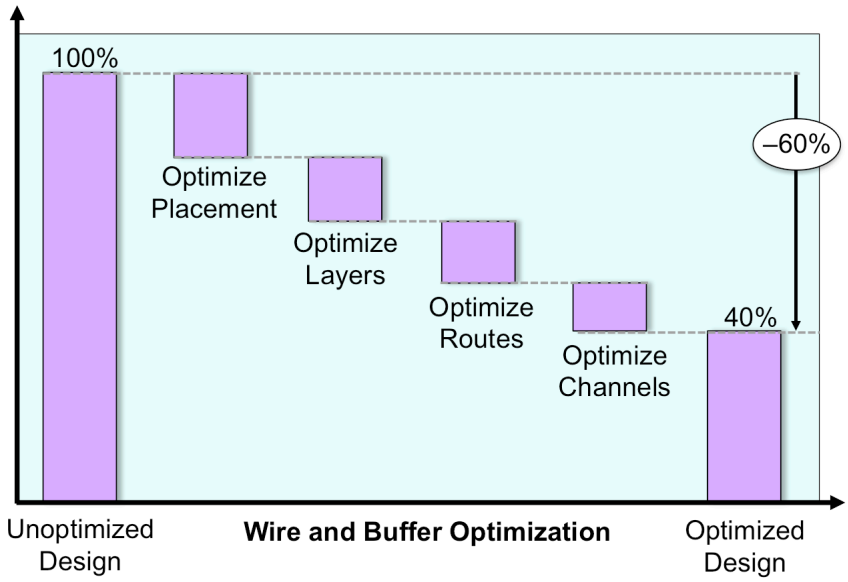

Figure 5. Optimizing Orion. Compared with AMBA AXI, NetSpeed's technology saves power by optimizing the interconnects, routers, FIFO buffers, and distributed arbitrators. (Source: NetSpeed)

Gemini Adds Cache Coherence

Orion was NetSpeed's first product. Gemini builds on it by adding directory-based cache coherence for multicore SoCs. Directories are more efficient than cache snooping for large multicore designs, but they require additional memory—typically, about the same amount as the sum of the L2 cache tags. The directories can grow quadratically with the number of clusters, so NetSpeed uses several patented techniques to keep the growth closer to linear. (Because the directories don't manage I/O coherence, the number of I/O agents doesn't affect the directories' size.)

Gemini can integrate up to 64 CPU clusters, GPUs, DSPs, or acceleration engines and up to 200 peripheral cores that are I/O coherent or noncoherent. NetSpeed has simulated and verified 64-cluster designs on a Cadence Palladium emulator. Both Orion and Gemini are correct-by-construction NoCs that eliminate protocol- and network-level deadlocks, no matter how complex the network.

NocStudio calculates the interconnect bandwidth required to ensure coherence among all the processing cores and I/O interfaces, then generates the appropriate number of coherence controllers. Each controller can handle about 32GB/s of coherent traffic at 1.0GHz. Many designs require only one controller, depending on the application. Memory bandwidth is a factor, because it limits the coherence bandwidth. Additional factors are the core count and cache sizes, although the controllers are not assigned to particular CPU clusters. Designs with multiple coherence controllers divide the address space among themselves so they can work in parallel to achieve higher bandwidths.

NocStudio generates additional elements, such as a coherent traffic accelerator that performs multiple directory lookups in parallel and ensures that transactions will complete in the correct order. Gemini also supports distributed virtual memory (DVM), which helps the operating system manage memory. When the OS updates a translation lookaside buffer (TLB), it must broadcast the invalidated entries to all agents using the page tables. Those agents include CPUs and virtualized I/O devices. ARM's AXI Coherency Extensions (ACE) can distribute snoops to various agents using DVM, so Gemini supports those protocols.

Gemini is scalable for different-size chips. NocStudio can generate relatively small NoCs for low-power mobile processors or much larger NoCs for the manycore processors in enterprise networking equipment. Like Orion, it supports NoCs with up to 8 physical networks and 32 virtual networks.

No matter which NoC the chip design employs, NocStudio's final output includes performance, power, and area statistics; the RTL files required to synthesize the NoC; a C++ functional model; and verification test benches. A separate fail-safe layer called the RegBus supports post-silicon tracing and debugging. NocStudio even generates English-language technical-reference manuals that document the chip's internal specifications and explain why the tool made various design decisions.

Bridging Architecture and RTL

As a small startup in a market crowded with technology startups, it's easier for NetSpeed to promote its more-familiar technology (licensable NoCs) than to claim a paradigm shift in a less-familiar technology (front-end chip design). The company could probably succeed on the strength of its NoCs alone. But NetSpeed wants to aim higher by bridging the difficult gap between architecture specification and design implementation.

All design automation imposes trade-offs. Software compilers generate binary code that's less efficient than expertly written assembly language; synthesis compilers generate circuit designs laden with NAND gates instead of smaller logic gates. The industry accepts these compromises to reduce labor costs and accelerate time to market. Likewise, NocStudio may generate a NoC topology that isn't the absolute best solution possible. But because it enables engineers to rapidly reconfigure the NoC and instantly get performance feedback, it encourages alternative approaches while removing much of the doubt associated with comparing different solutions. Also, its correct-by-construction methodology prevents the worst tapeout-killers, such as network deadlocks.

However, what happens if NocStudio's high-level floor plan doesn't match the chip's final place-and-route floor plan? Stretching and rerouting the wiring to reach the IP blocks' new locations could nullify the automated optimizations. If NocStudio reoptimizes the network to match the new floor plan, the NoC may be less efficient (in latency, bandwidth, wiring, or power consumption) than the original ideal configuration. But any NoC can suffer from such changes, no matter how it was designed. NetSpeed's user-directed automation is still an advantage because reoptimizing takes less time and is less prone to error.

NetSpeed's application of macroscale networking algorithms to chip design is novel and bears consideration. NoCs must accommodate a wide variety of IP cores with different characteristics, including multiple clock-frequency and voltage domains and varying memory requirements. Those kinds of low-level differences are invisible to macro networks. Can NocStudio's automation cope with such complexities across many different chip designs, applications, and use cases? NetSpeed says its tools account for those differences and that the networking principles embodied in NocStudio's algorithms have more to do with traffic congestion, network scalability, QoS, and deadlock avoidance.

Design Automation Accelerates Projects

NetSpeed has several satisfied customers, although they prefer to remain anonymous for now. One is a top-tier ASIC design team that is using NocStudio and Orion for a massively parallel processor with more than 1,000 cores. Previous generations of this product family used proprietary interconnects, but the growing complexity of the new design was getting out of hand. The architects evaluated and rejected other licensable NoCs because they could not meet the chip's demanding performance requirements.

In the past, the architects modeled the entire chip in SystemC, tested various use cases, and modified the NoC as necessary. Each design iteration took about a month. When the design seemed satisfactory, engineers ported the SystemC model to RTL and repeated the tests to verify that the RTL was correct and free of deadlocks. The RTL porting alone took six to nine months. Now, with NocStudio and Orion, the architects can model a NoC and automatically generate the RTL. "It's basically a network compiler," says the design team's project leader. "We are iterating the design in hours instead of in days or weeks. It has offloaded our architects to work on other things. And it's correct by design, with no deadlocks."

Another benefit is that the architects no longer must overprovision the NoC and hope it can handle every possible use case. Instead, they can model the NoC's performance accurately enough to achieve an optimal design. As the project leader points out, this capability becomes more important when targeting advanced IC processes, because the wire-routing metal layers are not keeping pace with the soaring transistor budgets. Optimizing the metallization can significantly reduce the chip's fabrication cost.

Front-end design automation is surely coming in some form or another. SoCs are becoming too large and complex for existing design methodologies to persist much longer. As was seen when software compilers and RTL synthesis compilers gradually won adoption, skilled-labor efficiency and time-to-market pressures eventually outweigh other considerations. Without more design automation, the complexity will become so daunting that fewer companies will risk designing their own chips—but the alternative is reverting to general-purpose processors that can't match an optimized solution. Sooner or later, the industry will embrace front-end design tools that inevitably will look very much like NocStudio. Architects who need a scalable, high-performance, correct-by-construction SoC interconnect should evaluate NetSpeed's technology, especially if the design requires cache coherence.

Tom R. Halfhill is a senior analyst at The Linley Group and a senior editor of Microprocessor Report. *The Linley Group offers the most comprehensive analysis of the microprocessor industry. We analyze not only the business strategy but also the internal technology. Our in-depth reports cover topics including embedded processors, mobile processors, network processors, base-station processors, and Ethernet chips. For more information, see our web site at www.linleygroup.com.*